3. Finding home on Twitter. By Eduard Campillo-Funollet

Introduction

The concept of *home* is present in everyday life in many different ways. Even from the limited point of view this personal experience provides, it is easy to find many different ways to think about *home*, what *home* is, and what *home* means. Therefore it is natural to ask how different concepts of *home* are used today.

In the digital age, a major space for communication is the Internet, and therefore an excellent framework to study how people are using a certain concept. The digital world has many characteristics that distinguishes it from other forms of communication, but it also has a significant advantage: it is easy to get huge amounts of data in a short time.

The context of the concept of *home* is of course related to a number of possible meanings; a basic example: during a baseball match, there is a high probability that one talks about the *home plate*, the final base that a player must touch to score. Despite the fact that the underlying meaning of the word *home* here is still "a place of origin", the actual use of the word is not linked to other concepts of home, such as the house where one lives.

Given the previous considerations, I will study how the concept of *home*, and in particular, the word *home* is used in one social network, Twitter. In contrast to other networks, all publications of Twitter users are publicly available to everyone and mainly in a text form. For these reasons, Twitter is an excellent source of data.

In performing this study, I have two different goals. First, as mentioned above, I aim to present some insight on what *home* means in the context of Twitter. In particular, I am interested in what can be learned about the meaning of the word *home* in the limited context of a tweet.

The second goal of this chapter is to showcase the tools that mathematics and data analysis provide for humanities and social sciences. The aim is not to perform in-depth analysis, but rather to demonstrate an approach, and elicit questions to experts in the field. Therefore, I have skipped technical details that do not have relevance in this context, but included elementary presentations of the techniques to provide a clearer idea of what I have done.

The data for the study was collected for a period of 24 hours, between May 30th and May 31st, 2017. The size of the sample is rather small - 250,000 tweets, reduced to about 120,000 after cleaning the data - and in particular it will capture transient phenomena related to events that happened on that particular date. There are two reasons to limit the size of the sample. Firstly, since the goals of the study are to demonstrate what can be done rather than find a new phenomenon, a small but still significant dataset is sufficient. Secondly, since this book was produced in a limited time frame; conse-

quently I planned a study that could be fully contained in that given time frame. The analysis was planned and prepared in advance, but the data collection and analysis were performed during the event at which this book was produced.

All the data collection and analysis was performed in Python. The tweets were collected using in-house code via the Twitter Streaming API (Twitter, 2017). The analysis used in-house implementations of some algorithms and the Natural Language Toolkit module (Bird, 2006). Figures were produced with the module *matplotlib* (Hunter, 2007) and *wordcloud*.

The chapter is organized as follows. After a brief description of Twitter as a social network, and how its characteristics impact this study, an explanation the analysis of the data as it unfolded follows. I begin with a preliminary exploration of the data, and then proceed to frequency analysis, a sentiment analysis and finally to the application of a word disambiguation algorithm. At each step, I explain my reasoning and observations, and how these lead to the next step.

Twitter: a micro blogging social network

Twitter is a popular social network, built on the concept of *micro blogging*. Users publish short texts, occasionally including other media. Each of these publications is called a *tweet*. The network was created in 2006, and grew quickly. Over three hundred million people worldwide are now active users of Twitter. On average, about 500 tweets are published every day (Twitter, 2017).

In comparison to other social networks, the main characteristic of Twitter is the limited length of the posts: a tweet is limited to a maximum of 140 characters. Originally, the limitation was imposed to allow tweets to be sent by SMS (Sarno, 2009), but it has since become a defining feature of Twitter. Media such as pictures and videos do not count towards the length limit, but any form of text, including hyperlinks, does.

Twitter encourages users to include keywords in their tweets using hashtags. A hashtag is simply a word - it may be several words written together, without spaces in between - with the character # at the beginning. A list of the most popular hashtags - the trending topics - is updated in real time, effectively inviting the users to use these hashtags in their publications.

In contrast with other social networks such as Facebook, all Twitter user profiles are public. Any internet user, whether a registered Twitter user or not, can read all the published tweets. A user must register to publish new tweets.

Twitter does not require any formal verification of identity to register. The user may remain anonymous if he or she wishes, or can use a pseudonym. Twitter provides a verification service that is used by important companies or public people, for example, celebrities and politicians. The vast majority of the accounts are not verified.

The data: tweets with the word*home*

My goal is to use established data analysis tools to identify and characterise the different concepts of *home* in Twitter. At this point, I have not assumed any knowledge on how the concept has been

used in Twitter. I started by analysing the language in the tweets containing the word *home*, either as a word in the text or as a hashtag.

Some tweets implicitly refer to *home* without actually using the word *home*. Finding these tweets is out of the scope of this study, but a characterisation of the tweets that explicitly use the concept *home* will inform any future study that targets implicit uses of the concept.

As mentioned, hashtags are often composed of more than one word, but written without spaces. For example, we can find tweets with hashtags like *#goinghome* or *#gohome*. I have chosen not to include these tweets as there are inherent limitations on the ways to filter them: I could use a handcrafted list of hashtags, but this would be subjective; I could find hashtags containing the characters *home*, but then I would find completely unrelated concepts, e.g *#homeopathy*. Therefore, I have chosen to limit myself to tweets that contain *home* as an isolated word, or as a hashtag (*#home*).

Since my approach is based on the use of the word *home*, all my data and conclusions are limited to English-speaking Twitter users. A similar approach could be used to analyse tweets in other languages, but it is out of the scope of this study.

The hashtags identify the keywords of the text. A tweet with the hashtag *#home* possibly has a stronger connection with the concept of *home* than a tweet that simply uses the word in the text. To incorporate this fact in the analysis, I studied both the full text and the set of hashtags of the tweets. Note that a tweet that contains the hashtag *#home* will always have the word home in the text, i.e. the hashtag is part of the text. The opposite is not true: a tweet can contain the word *home*, but not the hashtag *#home*.

I collected the data for the study over a period of 24 hours. Although the time frame was short, the number of tweets was significant (250,000). On the other hand, some features may be related to a current trend, for example, a trending topic or a live event, or may be related to the particular time, for example, day of the week or season.

Preliminary exploration

Before starting with the analysis, I did a preliminary exploration of the data. The goal was two-fold: firstly, I wanted to pick up any striking feature that might affect the later analysis; secondly, since most of my analysis was based on language features, I wanted to ensure that the captured tweets satisfied the basic characteristics of a language.

I began with a manual exploration of the data. I randomly selected a few tweets from the data and read them. The first observation was that I was still capturing tweets that were not related to the concept of home. This was due to what was probably the most extended use of the word *home* in the Internet: the homepage of a website, as many URLs contain the word *home*, for example, in the format *<domainname>.com/home/<page>.html*.

I removed the tweets that only contain the word *home* in a URL. It was straightforward to do so, since Twitter uses a URL-shortener to help users to comply with the 140 character limit. In the shortened URLs, the word *home* is never present. I therefore excluded tweets that did not contain the word *home* in the text after shortening the URLs, unless they explicitly had the hashtag *#home*.

Here are four random tweets from the dataset:

- *RT @JillDLawrence: .@CNN says Trump is home alone, stressed out, gaining weight, realizing job isn't good fit for him. https://t.co/MGftsHH (*Lawrence, 2017)
- *как вы оцениваете AVAST HOME* (Avast Software, 2017)
- *RT @sidebae: i'll marry the guy that makes me want to come home as much as my bed does. (Ya Girl, 2017)*
- *#RelianceJio Jio may launch home broadband plan at Rs 500/100GB in 100 cities via https://t.co/Ga70xDbe9A* (Singh, 2017)

After removing the tweets with the word home only in an URL, I significantly reduced the tweets in languages other than English. Some did still remain but I chose to keep them in the dataset. If there was a significant use of the word *home* in a foreign language, it appeared in the subsequent analysis.

It is clear even from this very limited sample that the word *home* is often used in business-related tweets, such as advertisements. It is important to bear this in mind, since the concept of home in the context of an advertisement is most probably connected to home as the house where one lives, either as part of a product name, for example, home broadband, or in the advertisement text.

Some tweets also contain words that are not part of the actual text. The tag *RT* is included at the beginning of a *re-tweet* (a tweet from another user that is reproduced in another user's timeline). Therefore, this tag will have a high frequency, but is not relevant for my analysis. Similarly, words starting with the character @ are mentions of other users.

To ensure that my sample of tweets were consistent with natural language, and therefore that I was able to apply standard tools to analyse it, I confirmed that the tweets follow Zipf's law.

Zipf's law is an empirical law on the statistics of a large sample of words. The law states that the frequency of any word is inversely proportional to its rank in the frequency table. Most natural languages satisfy Zipf's law, and in particular, Pak and Paroubek (2010) establishes that the tweets in general follow Zipf's law. Therefore, if the sample of tweets with the word *home* is large enough, and if these tweets are not essentially different from an average tweet, our sample would also satisfy Zipf's law (Manning, 1999).

To visualise if the sample satisfied Zipf's law, I plotted the rank versus the frequency in a logarithmic scale. The logarithm transforms the relation to a linear correlation, and therefore the data will follow an approximately straight line if it satisfies Zipf's law.

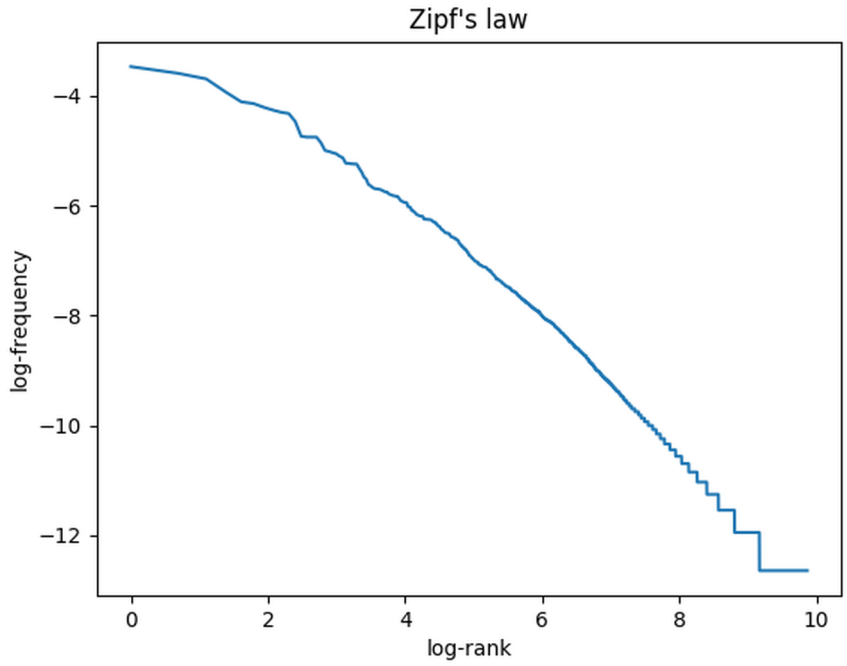

Fig. 2: A graph in log-log scale of the rank versus the frequency of the words in the data sample. Since the graph follows approximately a straight line, the data satisfies Zipf's law.

Because I selected tweets containing the word *home*, this word was over-represented in the sample, and I excluded it from the Zipf's law test. The rest of the words followed Zipf's law. I concluded that the language in the sample of tweets containing the word home was statistically natural, and so standard tools were suitable for the analysis.

Frequency analysis

The next step was to study the word frequency in detail. In particular, I used the frequency to get an initial idea of the *topics* in the tweets, but I was also interested to see if any word was present with a much higher frequency than others. The latter might already imply that the majority of tweets used *home* in the a similar way.

There were two facts to take into account before starting the frequency analysis. Firstly, I already knew that the word *home* has a high frequency: it is part of all the tweets in the sample. However, I did not exclude it from the analysis, because it mght be relevant to see if there were any other words with a similar frequency.

The second consideration is usual in analysis of natural language texts. Most of the languages had words with very high frequencies that were not relevant in this kind of analysis, and if these words were not excluded, they could mask important features of the rest of the words. These words are called stopwords. Examples of stopwords are pronouns (I, you, he, she...), common verbal forms (have, has, is, are), or very common words (under, up, too...). I used a standard list of common words provid-

ed in the package w*ordcloud*, and I also included the retweet tag *RT* as a stopword to exclude it from the analysis.

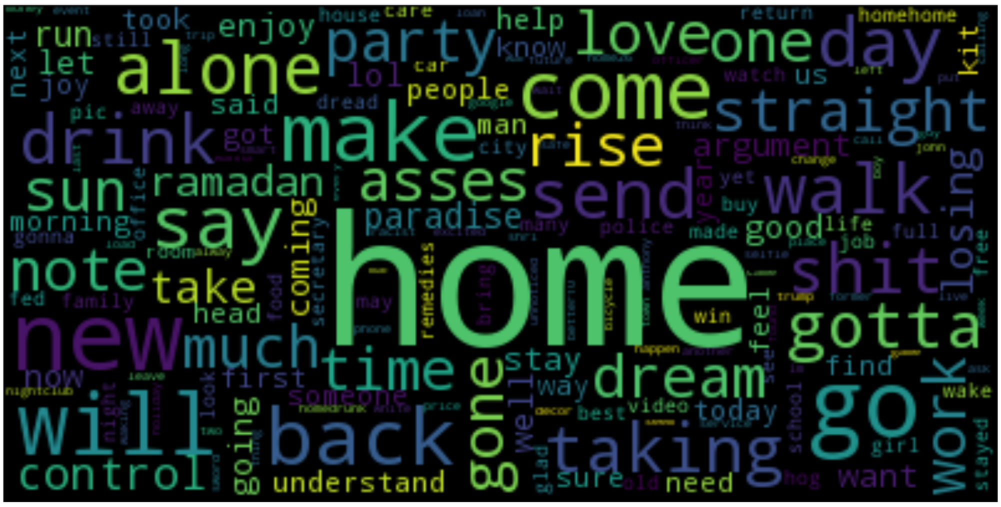

Fig. 3: A wordcloud of the most common words in the tweets that contain the word "home".

After the word *home*, the most common word in the tweets was *go;* this immediately suggested the frequent expression "go home", but at this point I did not have enough information to infer how it was used, and had to study the collocations of two words to get insight on this. Many other words in the word-cloud strongly suggested collocations, for instance "back home", "home alone" or "new home".

A few of the words in the word-cloud suggested routine actions and the contrast of home and work. For example, "morning", as well as words that are not included in the cloud such as "5:50", refer to the action of waking up. In both cases, *home* in this context was a reference to the physical place that one inhabits, the place where one sleeps. Other words related actions, for example, "drink", which are also possibly connected to same concept.

A final observation from the simple word-cloud was that most of the words had a positive or neutral connotation. For example, the words *party*, *love*, or *good* are usually associated with positive feelings. Only one clearly negative word was present in the forty most frequent words: *shit*. A sentiment analysis would provide more insight on this.

The same process was done for the hashtags, and the results differed. As expected, the words that were popular because of a current trend also had their own hashtag. The hashtags are the mechanism to find the tweets relevant for a user, and so if a topic is trending it will have its own hashtag.

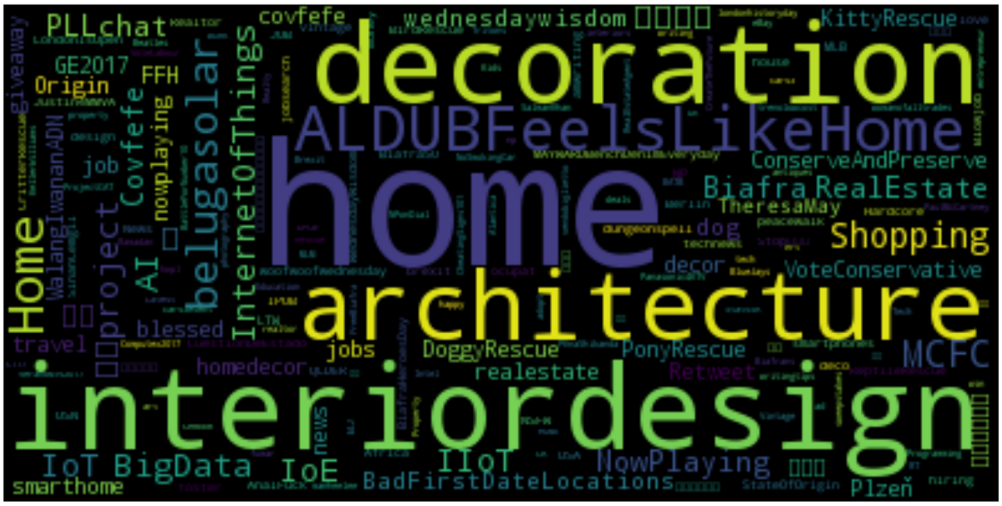

Fig. 4:

A word-cloud of the most common hashtags in the tweets with the word "home".

Since hashtags are promoted by Twitter to measure the *trending topics*, I expected some of the trending topics to appear in the sample too, even when the topic was not directly related to a particular concept of home. The best example of these was the hashtag #covfefe, a word used in a Tweet by the president of the United States, Donald Trump, on May 30th, 2017.

Words that I previously interpreted as coming from business users, like architecture, were more frequent as hashtags. Although this was not a confirmation that the actual tweet was business-related, it is consistent with the fact that a business carefully adds hashtags to its tweets to target potential customers. Another good example that did not arise in the first word-cloud, but was present here is #realestate.

The word "Biafra" deserves a special attention. The Republic of Biafra was a secessionist state in Nigeria, that existed for few years in the late 1960s. On May 31st, 2017, the leader of the Indigineous People of Biafra declared a *sit-at-home* protest. Hence, the tweets containing both the word Biafra and the expression *sit-at-home* were trending. Nevertheless, this is yet another use of the *home* a dwelling.

The hashtags provided a better insight into the theme of a tweet, and so I could extrapolate the context of the use of the word *home*. For instance, the hashtags #ponyrescue and #kittyrescue refer to animal rescues, usually finding a new home for an animal. Again, home is the dwelling, of an animal in this case.

I then considered all the bigrams. A bigram is a collocation of two words, i.e. two words that appear together in the text. The most frequent bigrams also contained the word home. Expressions like "go home", "back home" and "took home" were among the most frequent bigrams. In most of the cases,

these were expressions of movement, either of the individual or of an object, towards or away from home.

I found some of the words that were also in the first word cloud, such as *drink* or *party*.

In general, the sentiment expressions of the bigrams were still positive or neutral, with the exception of the bigrams containing *shit*. In any case, it was not possible to fully associate a positive or a negative feeling to expressions such as "shit taking" or "ain't shit" at this point.

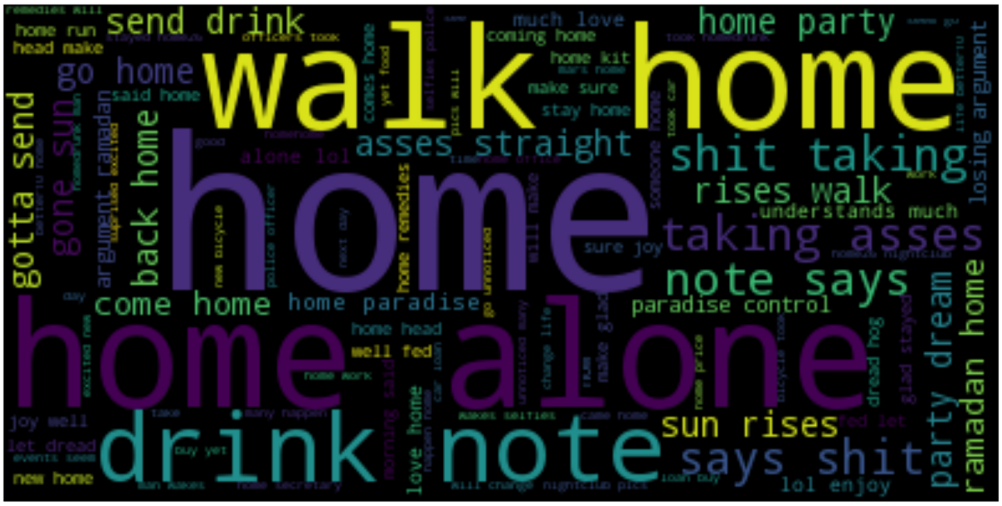

Fig. 5: A word-cloud of the most common bigrams in the tweets containing the word "home".

Sentiment analysis

Sentiment analysis is a popular tool to analyse the content of a text. The goal of sentiment analysis is to identify and classify the *sentiments* expressed in the text. In its simplest form, a sentiment analysis algorithm takes a text and outputs a value between 0 and 1 for positive, neutral, and negative sentiments. A value of 0 in one of the categories means that the type of sentiment is not present at all in the text. A value of 1 means that the text is clearly expressing the sentiment. Note that the category "neutral" includes all the cases that cannot be categorised as positive or negative: if it is not possible to say if a text is positive or negative, it will be classified as neutral. Furthermore, a text have an score for each category, for instance the tweet "I think I'll just go home after this, today is seriously tiring" (Qoni, 2017) has the scores positive 0.145, neutral 0.855, negative 0.0.

Nowadays, many popular sentiment analysis algorithms are based on supervised machine learning techniques. In machine learning, there is a training dataset with the correct categories assigned by some reliable method. In the case of sentiment analysis, the method could simply be classification by hand. The training dataset is then used to *train* the algorithm: the algorithm *learns* what kind of text

goes in each category. Once the algorithm is prepared, it uses the information from the training set to classify any new text.

Machine learning algorithms are powerful, but they have the drawback of requiring good training datasets. It was not clear how to obtain a training dataset for the tweets containing the word *home*. I could simply have used a training set based on all tweets, since many of such datasets are freely available, but the algorithm could become biased if the sentiments in the tweets containing the word *home* were expressed in a different way.

The alternative was to use rule-based algorithms. In a rule-based algorithm, a set of rules is used to obtain the final result. A rule is for example to assign points according to the presence of words from lexicons of positive and negative emotions, for example, love, nice or good for positive emotions; hurt, ugly or sad for negative emotions. The rule can be refined later, for instance with a second rule that updates the already positive or negative score if there are words like "very".

Rule-based algorithms are less specific than machine learning algorithms based on a good training set, but offer a more systematic classification. The fact that the algorithm user knows the rules that will be applied provides a deeper understanding of the final result.

To analyse the sentiments in the tweets containing the word *home*, I used VADER (Hutto and Gilbert, 2014), a rule-based algorithm that was designed specifically for social media text.

Firstly I looked at the average sentiment. I noted that already, in the frequency analysis, positive words were more abundant than negative ones. Furthermore, given that a tweet is a relatively short piece of text, most of the tweets would get a high neutral score. An average tweet scored over 0.8 in the neutral category. In other words, an average tweet cannot be clearly associated with a positive or negative feeling.

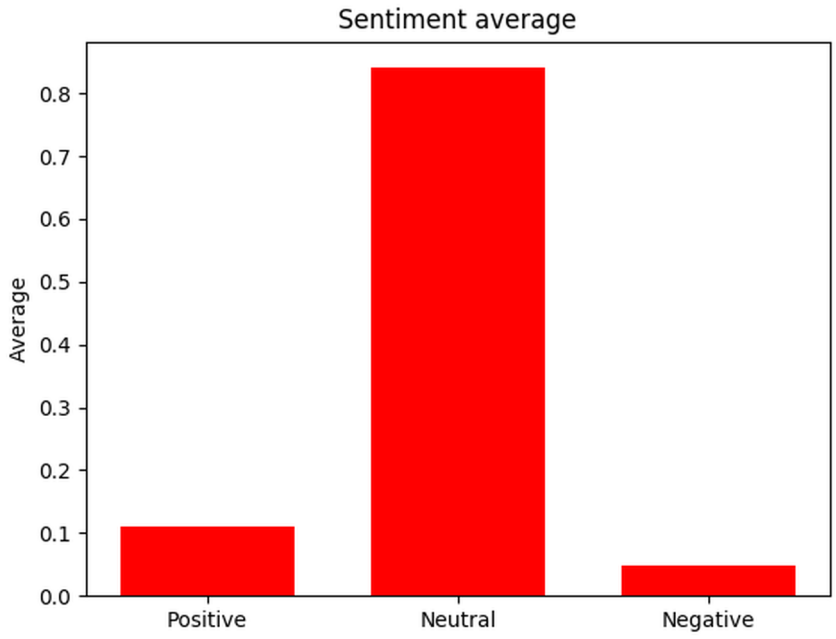

Fig. 6: The average sentiment score for the sample of tweets

The average score for positive was slightly over 0.1. This does not mean that there are not extremely positive tweets, but rather that given the high average of neutral scores, a typical tweet is mostly neutral with a pinch positive.

The negative score is an average about one half of the positive score. This confirmed the observation that I made from the frequency analysis: if I considered tweets that I could interpret clearly as positive or negative, I found many more positive texts than negative ones.

I then looked more closely at the sentiments. So far, I have discussed the average, but I could also look at what type of sentiment is predominant in each tweet. In other words, I would not look at the proportions between the three sentiments in each tweet, but rather, say, a tweet is for instance positive if the highest of the three scores - positive, neutral, negative - is the positive score.

I already knew that the neutral score is on average higher. Therefore, most of the tweets had a predominant neutral sentiment. More precisely, 99.2% of the tweets in the sample were essentially neutral. I studied the tweets with higher neutral score in detail, to see what concepts of home were used.

The tweet with the highest neutral score was "*[Help] Nfs crashes while pressing home on 10.1.1 jailbroken via /r/jailbreak https://t.co/tzFF5SKqFW*". This tweet refers to a problem with a computer game - Nfs is short for Need for Speed - that apparently stops working when the key *home* is pressed. This is a common meaning of *home* in the digital age: the "Home" key is found on most computer keyboards. Originally, the key was used to go to the top menu on non-graphical computer applications, the label "Home" in reference to the place where one begins. In modern computers the key is used to move the cursor to the beginning of a line or page. Although the reference to "Home" is not that clear - it is not natural to use "Home" to refer to the beginning of a line - the underlying concept is the same: move back to the place where one begins.

Other tweets with high neutral scores are questions. Unless the author is already looking to suggest an answer, a question is usually formulated in a neutral tone, and therefore there are not predominantly positive or negative words. I looked for instance at the tweet "*HOW HIRING HANDYMAN REDUCES HOME RENOVATION COST? https://t.co/eHZHLnZxRO*", which is essentially an advertisement of a home renovations company. Here *home* is used in the usual meaning of a physical house.

I then examined the negative tweets. The tweet with the highest negative score is "*Home alone):*". *Home* refers again to the dwelling, this time accompanied with the word "alone", that is tagged as predominantly negative in the sentiment analysis. But what gives a high negative score to the tweet is the emoticon "): "---a sad face, written in reverse order.

The sentiment analysis model VADER takes into account the most popular emoticons - smiling :-), grinning :-D, sad :-(and crying :_(- to assign the scores to each emotion. It is worth noting that the implementation of VADER used for this analysis does not consider the emojis, one character representations of the emoticons and other objects, extremely popular on mobile phones. Including emojis in the analysis might change the results. For instance, the tweet "*my brother coming home today*  ", is the second tweet with the highest neutral score, but the grinning face emoji at the end suggests a positive sentiment.

The highest positive sentiment tweets are also short and characterised by an emoticon. For instance, I observed tweets like "*Home =)*" and "*Home :)*", with positive scores over 0.7. Another common reference

is to safety, e.g "*Home safe*", "Thank God I made It Home Safe". I found again that home is used here to refer to the house that one inhabits.

In some positive sentiment tweets, I observed references to home that could be more subjective, but it is not possible to fully understand them in the context of a short text. For instance, the tweet "*Feeling like home*" is a reference to simply a place where one feels comfortable, safe. The author could be using the phrase to mean that he is in a very comfortable place, but could also be referring to some feeling that is reminiscent to his or her place of origin.

Word sense disambiguation

So far, I analysed the most frequent words in the tweets that contain the word home, and I used those words to infer the concepts of home that are in place in Twitter. I also performed a basic sentiment analysis of the tweets, to get an insight of the concepts of home used in tweets that clearly express a sentiment. In any case, this analysis is limited to the amount of tweets that I can study manually. My goal is to find if the concept of *home* is used with other meanings, but since the uses that I already mentioned are predominant, it is difficult to find these cases by simple observation.

To overcome the difficulty, I use a word sense disambiguation algorithm. Word sense disambiguation techniques are algorithmic tools used to infer the meaning of a word, among all possible meanings, given the context. Although these techniques are not completely reliable, they provide a systematic method to study the meaning of a word in a given text. A limitation of the use of these techniques in Twitter is tweets are short pieces: there is not much context to analyse. This must be take into account when considering the conclusions of the analysis.

As in the case of sentiment analysis, there are machine learning methods for word sense disambiguation. These methods have the same drawbacks as previously mentioned for the sentiment analysis. Therefore, I use an algorithm based on the occurrence of words in the surroundings; essentially, the method based on statistics.

The algorithm is known as Lesk algorithm, and it was introduced in 1986. The idea behind the algorithm is to consider a series of definitions of a word, as they will be in a dictionary. If a word is used in the sense of a particular definition, one could probably find words of the definition near the word that we are analysing. In other words, given a text, the algorithm the definition that have more words in common with the text (Lesk, 1986).

The Lesk algorithm is obviously limited by the definitions. Very succinct definitions offer less possibilities to find words in common. Furthermore, the same meaning of a word can be defined using very different vocabulary, and the algorithm is very sensitive to the exact words used in the definition. Although I only used a basic implementation of Lesk, there are many ways to extend the algorithm and overcome these difficulties. For example, one can use thesauruses and synonyms dictionaries to avoid being sensitive to the exact wording of a definition.

For the definitions, I use the lexical database provided by Wordnet (Miller, 1995). The following list includes the seventeen definitions of home that I use for the analysis:

1. Where you live at a particular time.

2. Housing that someone is living in.
3. The country or state or city where you live.
4. (Baseball) base consisting of a rubber slab where the batter stands; it must be touched by a base runner in order to score.
5. The place where you are stationed and from which missions start and end.
6. Place where something began and flourished.
7. An environment offering affection and security.
8. A social unit living together.
9. An institution where people are cared for.
10. Provide with, or send to, a home.
11. Return home accurately from a long distance.
12. Used of your own ground.
13. Relating to or being where one lives or where one's roots are.
14. Inside the country.
15. At or to or in the direction of one's home or family.
16. On or to the point aimed at.
17. To the fullest extent; to the heart.

I apply the Lesk algorithm, using the previous set of definitions, then the sample of tweets containing the word *home*. The previous observations based on word frequency and sentiment analysis suggest that there are one or two very predominant concepts, so I expect the algorithm to find that a few of the definitions are more frequent.

According to the Lesk algorithm 40% of the tweets use *home* in reference to definition 1, the place where one lives at a particular time. It is difficult to disintguish in a short text like a tweet if the exact reference is to the general area or to the house that one inhabits in particular. The algorithm naturally leans towards the more general definition, which in this case is definition 1 (in comparison to definition 2).

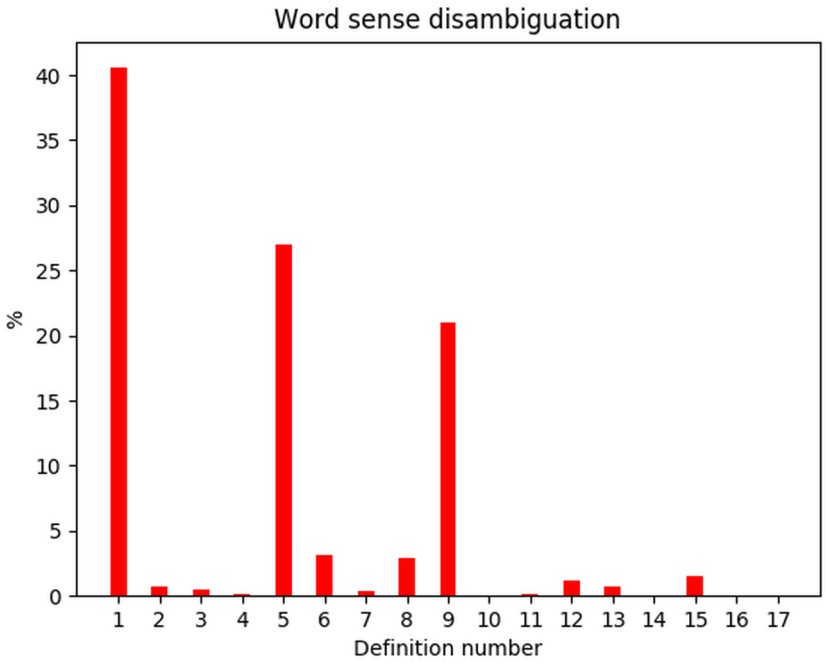

Fig. 7:

The proportion of tweets that are classified in each definition of *home* according to

the Lesk algorithm.

Note that definition 5, the second most frequent definition according to the algorithm, is also connected to the concept of "the place that one inhabits". The main difference between definition 1 and definition 5 is a sense of temporality: the place where one is stationed is home until the one is moved to another destination. This temporality is hard to grasp from a short text like a tweet.

The third most common definition is definition 9. Although this refers to an institution where people are cared for, the underlying concept is also the place where an individual stays. This definition is tagged in many of the tweets from animal rescue centres, e.g. the tweets with the hashtags #kittyrescue or #ponyrescue, but the use of the word *home* there is actually closer to definition 2. A typical tweet is "Find a new home for Missy #kittyrescue", with a picture of a cat attached. Clearly, the reference to home is to a house that one is living in, or even a house for the animal to live in, but not the the rescue centre as a home.

Overall, the top three definitions according to the Lesk algorithm are closely related and hard to distinguish in limited contexts. In contrast, they are clearly above other uses of the word home, for instance the algorithm did not find many references to the baseball *home*. The use of the word home in this context possibly increases during a baseball match.

Conclusion

The most common use of the word *home* on Twitter is related to the concepts of the house, or in general the place, where an individual lives. Although this is probably true in other contexts, when considering this conclusion it is important to keep in mind the public nature of Twitter. A published tweet is publicly available to any Internet user, and therefore an individual may not be sharing private ideas.

Even within the concept of *home* as a dwelling, some significant differences arise. The frequency analysis of the hashtags suggests that business-related Twitter accounts use the word *home* to publicise products and services, from decoration to architecture and construction. This suggest that to refine analysis, one should devise a methodology to distinguish business Twitter accounts from personal accounts.

The sentiment analysis suggests that on Twitter, the concept of home is more often connected to positive feelings than to negative feelings. Although the vast majority of tweets cannot be classified into strongly positive or negative, the number of strongly positive tweets is about three times the number of strongly negative tweets.

The word sense disambiguation algorithm confirms that the most common use of *home* on Twitter is in connection to the place where one lives. A finer distinction is difficult, given the limited context that a tweet provides. The conclusions can be improved by means of a more sophisticated word sense disambiguation algorithm, and also by using a more comprehensive datasets to include transient uses of the concept.

There are many possible ways to extend this study. I limited the data collection to words containing the word *home*, but a more comprehensive approach would include all the tweets with the character string *home*, even if it is part of longer word. During the data analysis, non-relevant uses of the words would have to be excluded.

Another natural extension will be to use the same approach in different languages, and in particular, to find the points in common on how the concept of home is used across different cultures. Even in the restricted context of Twitter, people from many different origins will have distinct ontologies on the concept of *home*.

Bibliography

Avast Software. (2017) *как вы оцениваете AVAST HOME.* (Twitter). 7 August. Available at: https://twitter.com/IosifKuneev/status/762227870911623168 (Accessed 31 May 2017).

Bird, S. (2006) 'NLTK: the natural language toolkit', *Proceedings of the COLING/ACL on Interactive presentation sessions*. Sydney: Association for Computational Linguistics. pp.69-72.

Hunter, J.D. (2007) 'Matplotlib: A 2D graphics environment', *Computing In Science & Engineering*, *9*(3), pp.90-95.

Hutto, C.J. and Gilbert, E., (2014) *Vader: A parsimonious rule-based model for sentiment analysis of social media text*. Available at: http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf (Accessed 2 June 2017).

Lawrence, J. (2017) *.@CNN says Trump is home alone, stressed out, gaining weight, realizing job isn't good fit for him.* (Twitter). 30 May. Available at: https://twitter.com/JillDLawrence/status/869712000745693185 (Accessed 31 May 2017).

Lesk, M. (1986) 'Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone', in *Proceedings of the 5th annual international conference on Systems documentation* (pp. 24-26). ACM.

Manning, C.D.; Schütze, H., (1999). *Foundations of statistical natural language processing.* Cambridge, MA: The MIT Press.

Miller, G.A. (1995) 'WordNet: a lexical database for English', *Communications of the ACM*, 38(11), pp. 39-41.

Pak, A., Paroubek, P. (2010) 'Twitter as a Corpus for Sentiment Analysis and Opinion Mining', *International conference on language resources and evaluation, proceedings of a conference,* Université Paris-Sud, Paris, pp. 1320-1326.

Qoni, P. (2017) *I think I'll just go home after this, today is seriously tiring*. (Twitter). 31 May. Available at: https://twitter.com/IamQonie/status/869865203197804544 (Accessed 31 May 2017).

Sarno, D. (2009) Twitter creator Jack Dorsey illuminates the site's founding document. Part I. Available at: http://latimesblogs.latimes.com/technology/2009/02/twitter-creator.html (Accessed 2 June 2017).

Singh, H. (2017) *#RelianceJio Jio may launch home broadband plan at Rs 500/100GB in 100 cities* via https://t.co/Ga70xDbe9A. (Twitter). 31 May. Available at: https://twitter.com/harendrasingh/status/869873185629495296 (Accessed 31 May 2017).

Twitter (2017) Available at: http://about.twitter.com (Accessed 2 June 2017).

Twitter (2017) Available at: https://dev.twitter.com/streaming/overview (Accessed 2 June 2017).

Ya Girl. (2017) *RT @sidebae: i'll marry the guy that makes me want to come home as much as my bed does.* (Twitter) 26 May. Available at: https://twitter.com/sidebae/status/868315770039803905 (Accessed on: 31 May 2017).

DOI: 10.20919/9780995786226/3